Chromosome-scale genome assembly provides insights into rye biology, evolution and agronomic potential

Rabanus-Wallace, M Timothy ; Hackauf, Bernd ; Mascher, Martin ; Lux, Thomas ; Wicker, Thomas ; Gundlach, Heidrun ; Baez, Mariana ; Houben, Andreas ; et al ; Praz, Coraline R ; Keller, Beat

Abstract: Rye (Secale cerealeL.) is an exceptionally climate-resilient cereal crop, used extensively to produce improved wheat varieties via introgressive hybridization and possessing the entire repertoire of genes necessary to enable hybrid breeding. Rye is allogamous and only recently domesticated, thus giving cultivated ryes access to a diverse and exploitable wild gene pool. To further enhance the agronomic potential of rye, we produced a chromosome-scale annotated assembly of the 7.9-gigabase rye genome and extensively validated its quality by using a suite of molecular genetic resources. We demonstrate applications of this resource with a broad range of investigations. We present findings on cultivated rye's incomplete genetic isolation from wild relatives, mechanisms of genome structural evolution, pathogen resistance, low-temperature tolerance, fertility control systems for hybrid breeding and the yield benefits of rye–wheat introgressions.

DOI: https://doi.org/10.1038/s41588-021-00807-0

ARTICLES
https://doi.org/10.1038/s41588-021-00807-0
nature genetics

Check for updates

OPEN

Chromosome-scale genome assembly provides insights into rye biology, evolution and agronomic potential

M. Timothy Rabanus-Wallace[1], Bernd Hackauf[2], Martin Mascher[1], Thomas Lux[3], Thomas Wicker[4], Heidrun Gundlach[3], Mariana Baez[5], Andreas Houben[1], Klaus F. X. Mayer[3,6], Liangliang Guo[7], Jesse Poland[7], Curtis J. Pozniak[8], Sean Walkowiak[8,9], Joanna Melonek[10], Coraline R. Praz[4], Mona Schreiber[1], Hikmet Budak[11], Matthias Heuberger[12], Burkhard Steuernagel[13], Brande Wulff[13], Andreas Börner[1], Brook Byrns[8], Jana Čížková[14], D. Brian Fowler[8], Allan Fritz[7], Axel Himmelbach[1], Gemy Kaithakottil[15], Jens Keilwagen[16], Beat Keller[4], David Konkin[17], Jamie Larsen[18], Qiang Li[19], Beata Myśków[20], Sudharsan Padmarasu[1], Nidhi Rawat[21], Uğur Sesiz[22], Sezgi Biyiklioglu-Kaya[23], Andy Sharpe[8], Hana Šimková[14], Ian Small[10], David Swarbreck[15], Helena Toegelová[14], Natalia Tsvetkova[24], Anatoly V. Voylokov[25], Jan Vrána[14], Eva Bauer[26], Hanna Bolibok-Bragoszewska[27], Jaroslav Doležel[14], Anthony Hall[15], Jizeng Jia[28], Viktor Korzun[29], André Laroche[30], Xue-Feng Ma[31], Frank Ordon[32], Hakan Özkan[22], Monika Rakoczy-Trojanowska[27], Uwe Scholz[1], Alan H. Schulman[33,34], Dörthe Siekmann[35], Stefan Stojałowski[20], Vijay K. Tiwari[21], Manuel Spannagl[3] and Nils Stein[1,36] ✉

Rye (*Secale cereale* L.) is an exceptionally climate-resilient cereal crop, used extensively to produce improved wheat varieties via introgressive hybridization and possessing the entire repertoire of genes necessary to enable hybrid breeding. Rye is allogamous and only recently domesticated, thus giving cultivated ryes access to a diverse and exploitable wild gene pool. To further enhance the agronomic potential of rye, we produced a chromosome-scale annotated assembly of the 7.9-gigabase rye genome and extensively validated its quality by using a suite of molecular genetic resources. We demonstrate applications of this resource with a broad range of investigations. We present findings on cultivated rye's incomplete genetic isolation from wild relatives, mechanisms of genome structural evolution, pathogen resistance, low-temperature tolerance, fertility control systems for hybrid breeding and the yield benefits of rye–wheat introgressions.

Rye (*Secale cereale* L.), a member of the grass tribe Triticeae and close relative of wheat (*Triticum aestivum* L.) and barley (*Hordeum vulgare* L.), is grown primarily for human consumption and animal feed. Rye is uniquely stress tolerant (biotic and abiotic) and thus shows high yield potential under marginal conditions. This makes rye an important crop along the northern boreal-hemiboreal belt, a climatic zone predicted to expand considerably in Eurasia and North America with anthropogenic global

[1]Leibniz Institute of Plant Genetics and Crop Plant Research (IPK), Seeland, Germany. [2]Institute for Breeding Research on Agricultural Crops, Julius Kühn-Institut, Sanitz, Germany. [3]Plant Genome and Systems Biology (PGSB), Helmholtz-Center Munich, Neuherberg, Germany. [4]University of Zürich, Zurich, Switzerland. [5]Federal University of Pernambuco, Pernambuco, Brazil. [6]Technical University Munich, Munich, Germany. [7]Kansas State University, Manhattan, KS, USA. [8]University of Saskatchewan, Saskatoon, Saskatchewan, Canada. [9]Canadian Grain Commission, Winnipeg, Manitoba, Canada. [10]The University of Western Australia, Crawley, Western Australia, Australia. [11]Montana BioAg Inc, Durham, NC, USA. [12]ETH Zürich, Zürich, Switzerland. [13]John Innes Centre, Norwich, UK. [14]Institute of Experimental Botany of the Czech Academy of Sciences, Olomouc, Czech Republic. [15]Earlham Institute, Norwich, UK. [16]Institute for Biosafety in Plant Biotechnology, Julius Kühn-Institut, Quedlinburg, Germany. [17]Aquatic and Crop Resource Development, National Research Council, Saskatoon, Saskatchewan, Canada. [18]Harrow Research and Development Centre, Agriculture and Agri-Food Canada, Saskatoon, Saskatchewan, Canada. [19]Huazhong Agricultural University, Wuhan, China. [20]West Pomeranian University of Technology Szczecin, Szczecin, Poland. [21]University of Maryland, College Park, MD, USA. [22]University of Cukurova, Cukurova, Turkey. [23]Sabanci University, Tuzla, Turkey. [24]Saint Petersburg State University, Saint Petersburg, Russia. [25]Vavilov Institute of General Genetics, Russian Academy of Sciences, Saint Petersburg, Russia. [26]TUM School of Life Sciences Weihenstephan, Technical University of Munich, Freising, Germany. [27]Warsaw University of Life Sciences-SGGW, Warsaw, Poland. [28]Chinese Academy of Agricultural Sciences (CAAS), Beijing, China. [29]KWS SAAT SE & Co, Einbeck, Germany. [30]Lethbridge Research and Development Centre, Agriculture and Agri-Food Canada, Lethbridge, Alberta, Canada. [31]Noble Research Institute, Ardmore, OK, USA. [32]Institute for Resistance Research and Stress Tolerance, Julius-Kühn Institute, Quedlinburg, Germany. [33]Production Systems, Natural Resources Institute Finland (LUKE), Helsinki, Finland. [34]Institute of Biotechnology and Viikki Plant Science Centre, University of Helsinki, Helsinki, Finland. [35]HYBRO Saatzucht GmbH & Co. KG, Isernhagen, Germany. [36]Georg-August-Universität Göttingen, Göttingen, Germany. ✉e-mail: stein@ipk-gatersleben.de

warming[1]. Currently, rye is produced on 4.1 million ha (http://www.fao.org/faostat/en/, accessed June 2020), 81% of which is in north-eastern Europe. More importantly, however, rye chromatin is commonly introgressed into bread wheat varieties to improve yield and thus rye genetic material is present in a far greater proportion of cultivated land area[2–5]. Rye is a diploid with a large genome (~7–8 gigabases, Gb)[6], 50% larger than the syntenic diploid barley and bread wheat subgenomes[7]. Like barley and wheat, rye entered the genomics era very recently. A virtual gene-order was released in 2013[8] and a shotgun de novo genome survey of the same line became available in 2017[9]. Both resources have been rapidly adopted by researchers and breeders[10–12] but cannot offer equivalent opportunities to the high-quality genome assemblies available for other Triticeae species[7,13–17].

We report a short-read based chromosome-scale genome assembly for rye inbred line 'Lo7' and demonstrate the potential of this new genomic resource by dissecting the incomplete genetic isolation of rye from wild relatives. We showcase detailed analyses of the genomic organization and complexity of gene families implicated in stress tolerance and pollen fertility. This resource will guide future rye breeding and provide immediate benefit in managing the trade-offs of using rye as a genetic resource in wheat crop improvement.

Results

Genome assembly, validation and annotation. We de novo assembled scaffolds representing 6.74 Gb of the estimated 7.9 Gb 'Lo7' genome from >1.8 Tb of short-read sequence (Methods; Supplementary Table 1 and Supplementary Note). These scaffolds were ordered, oriented and curated using: (1) chromosome-specific shotgun (CSS) reads[8], (2) 10x Chromium linked reads, (3) genetic map markers[9], (4) three-dimensional chromosome conformation capture sequencing (Hi-C)[18] and (5) a Bionano optical genome map (Supplementary Tables 2–7). After intensive manual curation (Supplementary Note), 92% of this assembled sequence (~78% of the estimated genome size) was arranged first into super-scaffolds (N50 > 29 megabases, Mb) and then into pseudomolecules. Shotgun reads (~947 Gb of data, ~120× mean depth-of coverage) were mapped back to the assembly to confirm a near-unimodal coverage distribution consistent with a high-quality assembly (Supplementary Table 8 and Supplementary Note). De novo annotation (Methods; Supplementary Table 9) yielded 34,441 high-confidence (HC) genes, including 96.4% of the BUSCO (v.3) near-universal single-copy ortholog set (Supplementary Table 1), 19,456 full-length DNA long terminal repeat (LTR) retrotransposons (LTR-RTs) from six transposon families (Supplementary Table 10)[19], 13,238 putative microRNAs (miRNAs) in 90 miRNA families (Supplementary Tables 11–17) and 1,382,323 tandem repeat arrays (Supplementary Tables 18 and 19). Full-length LTR-RTs represent a similar proportion of the total assembly in relation to genome size as shown by other recent Triticeae chromosome-scale assemblies (Supplementary Note and Supplementary Table 20) providing further evidence for high assembly quality and completeness[20]. Fluorescence in situ hybridization (FISH) to mitotic rye chromosomes confirmed agreement between in silico predicted and true physical distribution of distinct low- and high-copy probe sequences (Methods; Supplementary Note and Supplementary Table 21).

The rye genome follows similar organization as previously reported for other Triticeae genomes[7,13] (Fig. 1 and Supplementary Note): chromosomes are lacking recombination over ~50% of their physical length (Fig. 1a) and gene density increases by a factor of >10 towards the telomeres (Fig. 1b).

Gene collinearity plots (Fig. 1c and Supplementary Notes) between rye ('Lo7'), barley (cv. 'Morex') and wheat (cv. 'Chinese Spring'), confirm, with the exception of the gene-scarce zones surrounding centromeres, extensive genome collinearity. Genome expansion occurred rather uniformly over most of the chromosome arms with some acceleration toward distal regions, reflected by collinearity plots curving towards the telomeres (Supplementary Note). This expansion might be attributed predominantly to activity of LTR-RT families affecting the intergenic space[21,22]. We therefore estimated the time of highest insertion activity for the most frequent rye LTR-RT families RLG_SABRINA, RLG_WHAM and RLC_ANGELA (Methods; Fig. 1d–g and Supplementary Note). RLC_ANGELA elements did recently target this genomic niche and older RLG_SABRINA and RLG_WHAM expansions affected more proximal parts of the chromosome arms. Two distal regions on the long arms of rye chromosomes 4R and 6R, however, differed by a lack of the more ancient activity of the RLG_SABRINA and RLG_WHAM families, possibly highlighting regions affected by ancient translocation events from a rye with a different retrotransposon landscape to 'Lo7' (Supplementary Note). RLG_CEREBA elements (Fig. 1g) were active in centromeres, acting more constantly over longer time scales than the other frequent LTR-RT families.

Rye genome evolution. *Large structural variations—mechanisms of genetic isolation.* Megabase-scale inversions are a common feature of structural variation (SV) in the related barley genome[23]. In the absence of multiple rye genome assemblies, we sought to make a first survey of large SV prevalence among rye cultivars and wild relatives using three-dimensional conformation capture sequencing (Hi-C; Methods; Supplementary Note)[23]. In the comparison between two cereal rye cultivars 'Lo7' and 'Lo225', representing the two distinct heterotic gene pools in hybrid-rye breeding, megabase-scale inversions are apparent on four of the seven rye chromosomes (Fig. 2a and Supplementary Note). Among them, a 50-Mb inversion (comprising 382 HC genes) on chromosome 5R (positions ~650–700 Mb), coincides with a region lacking genetic recombination (Fig. 2b), providing genetic corroboration for its presence. This observation points to a previously undocumented source of unwanted linkage drag potentially affecting rye breeding efforts. Large inversions between 'Lo7' and other *Secale* representatives in the sample increase in number dependent on the phylogenetic distance to *S. cereale* and occur preferentially in the pericentromeric low-collinearity regions ($P < 0.001$, one-tailed empirical distribution derived from 10,000 simulations; Supplementary Note). Large SVs therefore provide a potential mechanism for localized collinearity loss between the pericentromeric regions of Triticeae species. This collinearity loss provides, in turn, at least one mechanism for effective genetic isolation during speciation[24].

Reticulate evolution of rye. Rye's divergence from its close relatives wheat and barley has not been comprehensively resolved. Using a draft assembly, Martis et al.[8] interpreted variation in sequence identity between rye and barley along chromosomes as a possible indicator of ancient species hybridization creating a 'mosaic' genome. Genome-wide estimations of fixation indices (F_{st}) and ABBA-BABA-based *D*-statistics have reflected varying levels of recent genetic exchange among *Secale* groups[25]. We confirmed, using *D*-statistics, that directional gene flow has occurred between rye groups (Methods; Supplementary Table 22). We then extended Martis et al.[8] sequence identity approaches with higher resolution proxies measurable across the chromosomes. Reticulation reduces the evolutionary distance between individuals of each species at any site where chromatin was secondarily exchanged, causing phylogenetic discordance among loci. Using the sequence identity of reciprocal best BLAST matches between rye CDS sequences and CDS sequences from barley cv. 'Morex'[13] and wheat cv. 'Chinese Spring'[7](Methods; Supplementary Note), we found no strong evidence of discordance among these genera: rye is more closely related to the bread wheat genomes than to barley across the whole genome (Table 1).

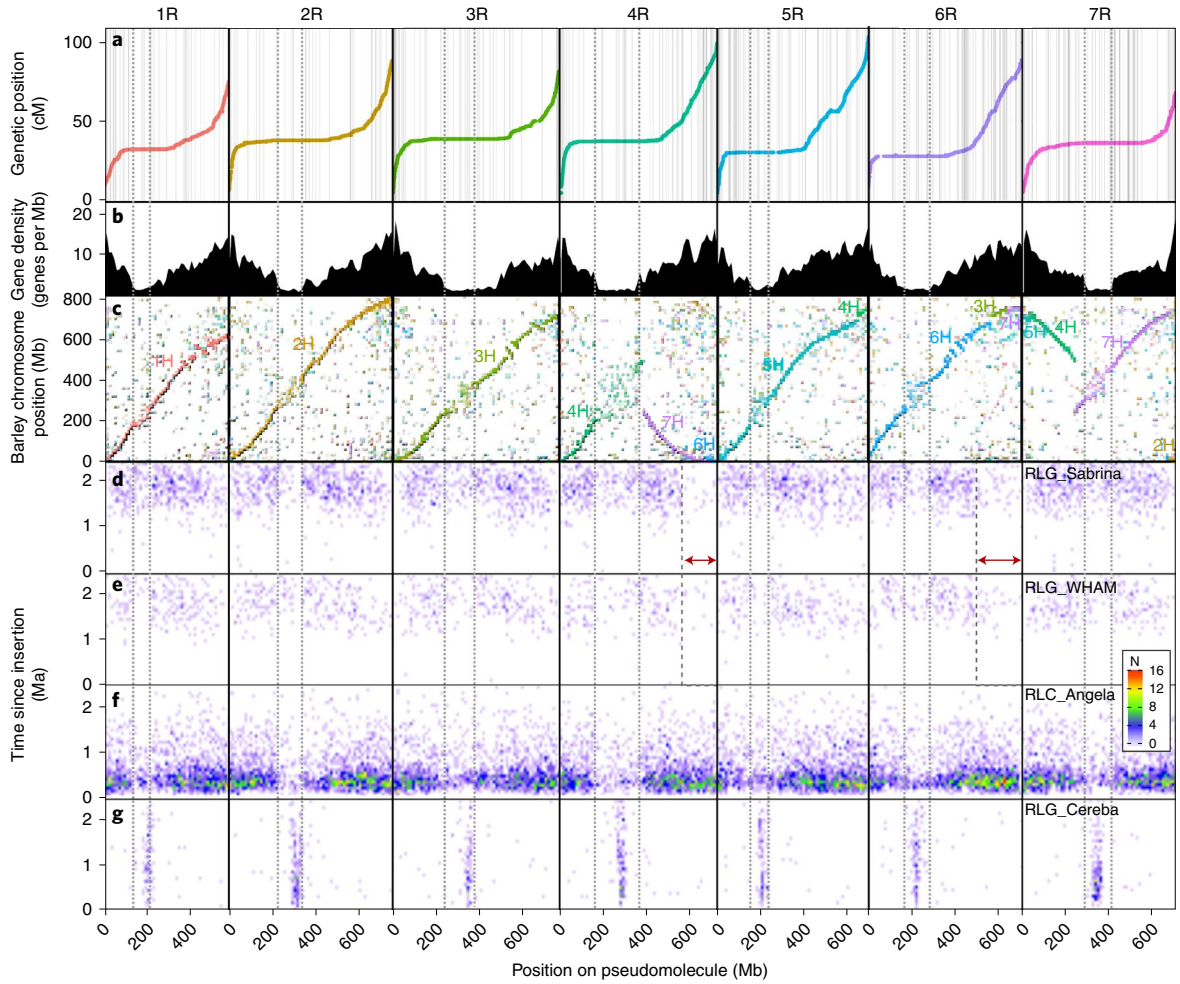

Fig. 1 | Rye ('Lo7') genome composition and structure over chromosomes 1R to 7R. Twin vertical gray lines in each chromosome denote the boundaries of the pericentromeric low-collinearity regions for each chromosome. **a**, Genetic map positions of markers used in assembly. Scaffold boundaries marked by gray vertical lines. **b**, Density of annotated gene models. **c**, Gene collinearity with barley (cv. 'Morex'), with the position on the 'Morex' pseudomolecules on the vertical axis. Text and point colors represent barley chromosomes as labeled. **d**–**g**, Positions and ages of four LTR retrotransposon families RLG-Sabrina (**d**), RLG-WHAM (**e**), RLC-Angela (**f**) and RLG_Cereba (**g**) in the genome, represented as a heatmap. Binned ages are on the vertical axis (from 0 million years ago, Ma, at the bottom) and bin positions are across the horizontal. Heat represents the number of TEs in each age/position bin (see legend inset). Red arrows mark notable changes in LTR-RT profiles.

We then produced an analogous analysis for *Secale* species by calculating identity-by-state (IBS) statistics between 'Lo7' and sequence data from a population of 955 cultivated and wild ryes (dataset of Schreiber et al.[25], expanded here by a further 352 genotypes; Methods). We used *k*-means clustering to define seven rye genetic clusters (Fig. 2c,d). In contrast to the intergenus comparisons, recent reticulation among rye clusters was strongly supported. In general, 'Lo7' is most closely related to *S. cereale* and *S. vavilovii*-dominated clusters and successively less related to *S. strictum*-like clusters and most distant from the *S. sylvestre*-dominated cluster (Fig. 2d; Supplementary Note). However, clear departure from this pattern occurs frequently (Fig. 2e,f and Supplementary Note). For example, at regions on chromosomes 1R and 4R (marked on Fig. 2e,f), *S. sylvestre*-like individuals are closely related to 'Lo7', often more closely even than some *S. strictum*-like individuals, suggesting recent genetic exchange between *S. cereale*-like and *S. sylvestre*-like genotypes. Pairwise F_{st} was calculated to assess the proportions of genetic variability within and between cluster groups and shows considerable variability across the chromosomes, especially comparing within- and between-group variability among *S. strictum*-like and 'domesticated'-like clusters (Fig. 2d–f). On chromosome 4R, for instance, F_{st} is almost 0.8 along the pericentromeric region (~200–400 Mb) but approaches zero at two interstitial positions (~600 and 720 Mb), corroborating incomplete genetic separation between these subgroups.

To investigate the effects of incomplete genetic isolation on recent selection pressures exerted on domesticated rye, we examined the ratio of nonsynonymous to synonymous mutations in exonic single nucleotide polymorphisms (SNPs) segregating among ryes (P_n/P_s) but which shared an allele with the consensus state of the three bread wheat genomes (a proxy for the ancestral state), thus surveying primarily recent mutations within rye lineages (Methods). Under equivalent selective regimes, P_n/P_s values for wild and domesticated ryes are expected to be approximately equal but we observed P_n/P_s divergence in the low-collinearity pericentromeric regions of chromosomes 1R, 3R and 4R. This divergence probably reflects the reduced efficacy of selection in such regions where recombination is limited (for example, by SVs).

On the basis of these collected observations, we concur with the Martis et al.[8] hypothesis that the genome of cereal rye is a mosaic in the sense that different rye species are not completely reproductively

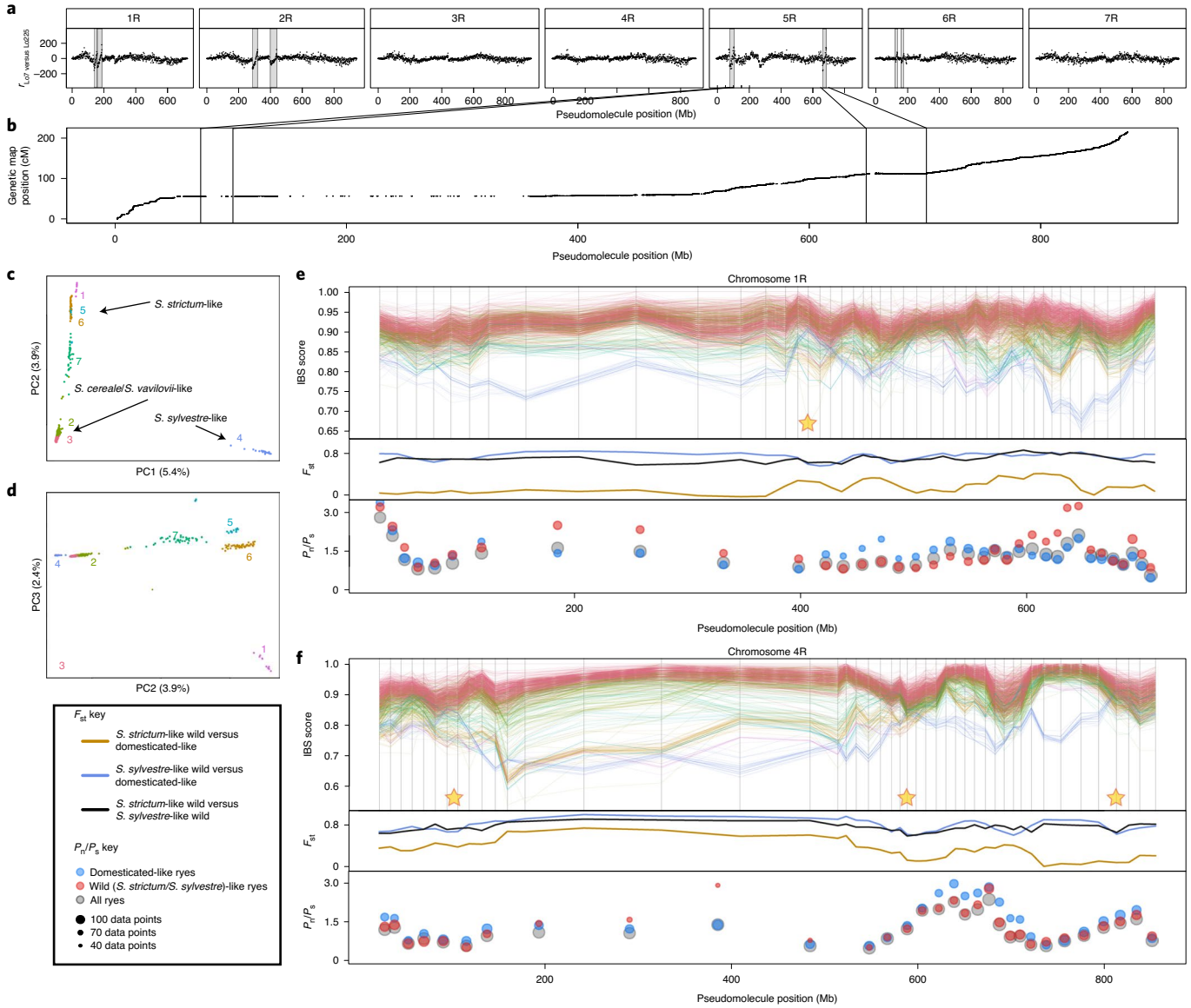

Fig. 2 | Dissecting the relationships among rye genotypes. a, Hi-C asymmetry detects SVs between the reference genotype 'Lo7' and *S. cereale* 'Lo225'. SVs result in discontinuities in *r*, the ratio of Hi-C links mapping left:right relative to 'Lo7'. Large inversions (marked) typically produce clean, diagonal lines. Visually identified candidate SVs are shaded. **b**, Detail of 5R genetic map marker positions showing how recombination rate relates to candidate SVs. The rightmost inversion marked on 5R corresponds to a region of suppressed recombination on chromosome 5R. The effect of other 'Lo7' versus 'Lo225' SVs on recombination was harder to confirm since they fall in already-low-recombining regions. **c**,**d**, PCA plots showing the relationships among genetically determined rye clusters for PCs 1 and 2 (**c**) and for PCs 2 and 3 (**d**). **e**,**f**, Binwise IBS, F_{st} and P_n/P_s statistics calculated across the chromosomes using the expanded Schreiber et al.[25] rye diversity panel data mapped to the 'Lo7' assembly. Exemplary instances are shown: chromosomes 1R (**e**) and 4R (**f**). The position of each bin on the genome is the mean pseudochromosome position of identified variable sites within that bin. Upper in each pane: binwise IBS scores of the panel genotypes compared with 'Lo7', with features discussed in the text marked with asterisks. Colors correspond to **d**. Middle in each pane: binwise F_{st} showing changes in genetic variance partitioning among and between subgroups across the chromosome. Line colors, F_{st} key. Lower in each pane: binwise P_n/P_s ratios (shown when $P_n + P_s >= 10$ for a given bin) for recently acquired rye polymorphisms (wheat outgroup). The values were calculated separately for different groups of ryes ('domesticated' (cluster 3) versus 'wild' (clusters 1,4–7)—see **d**) to allow detection of possible recent selective events affecting different rye groups. Point colors/sizes, P_n/P_s key.

isolated; however, we did not produce any evidence to suggest that the mosaic involves intergeneric hybridization.

Tracking the fate of rye chromatin in wheat improvement. The transfer of rye genetic material into bread wheat can provide substantive yield and stress tolerance benefits[26], though at the expense of bread-making quality[27]. These transfers involved a single 1BL.1RS Robertsonian translocation originating from cv. 'Kavkaz' and a single 1AL.1RS translocation from cv. 'Amigo' (Fig. 3)[3,4]. Due to the trade-off between yield and quality, wheat breeders must screen their programs for rye introgressions. Taking advantage of the new rye assembly, we implemented a high-throughput sequencing-based approach on four expansive wheat germplasm panels (Kansas State University (KSU), United States Department of Agriculture Regional Performance Nursery (USDA-RPN), International Maize and Wheat Improvement Center (CIMMYT), Wheat and Barley

Table 1 | Genome assembly and annotation statistics

Assembly	Raw scaffolds (after chimera breaking)	In chromosome-scale pseudomolecules	
Scaffolds	109,776	476	
Total length (Mb)	6,670.03	6,206.74	
N50 length (Mb)	15.16	29.44	
Length with chromosome assignment (%)	95.3%	100%	
Optical genome map			
Maps	5,601		
Total length (Mb)	6,660.18		
N50 length (Mb)	1.671		
Assembly/optical map alignment			
Total aligned length (Mb)	6,248.60		
Uniquely aligned length (Mb)	6,029.11		
Gene feature annotation	**High-confidence (HC) set**	**Low-confidence (LC) set**	
Number of genes	34,441	22,781	
Mean gene length	2,892	946	
Mean exons per gene	4.42	1.79	
Proportion of complete BUSCO set	96.4%	5.8%	
LTR-RT annotation	**Superfamily**	**Full-length copies**	**Mean age (Ma)**
RLC_Angela	Copia	11,128	0.53
RLG_Cereba	Gypsy	934	1.24
RLG_Sabrina	Gypsy	3,996	2.10
RLG_WHAM	Gypsy	1,457	2.06
DTC_Clifford	CACTA	1,480	NA
DTC_Conan	CACTA	516	NA
RLC total	NA	13,124	NA
RLG total	NA	1,973	NA
LTR-RT total	NA	15,097	NA

BUSCO, benchmarking universal single-copy orthologs (v.3; https://busco.ezlab.org/); NA, not applicable.

Legacy for Breeding Improvement (WHEALBI); Methods) segregating for both 1AL.1RS and 1BL.1RS. Translocations can be observed as obvious changes in normalized read depth across both the translocated and replaced chromosomal regions (Fig. 3b and Supplementary Note).

Human classification of a whole panel of karyotypes is still costly in terms of time. To alleviate this bottleneck, we developed an automated support vector machine (SVM) classifier that replicates human assignment with over 97% accuracy (Methods; Fig. 3c,d). We then demonstrated that the automated classifications predict yield. A mixed-effects linear model applied to yield data available for the autoclassified individuals in the KSU ($n = 19,677$) and USDA ($n = 29,035$) breeding panels showed that 1R introgressions could increase average yields up to ~4.55% (Table 2; Methods; Supplementary Tables 23–25). The 1AL.1RS karyotype significantly outyielded 1BL.1RS in the KSU panel but the reverse was true of the USDA panel (Table 2). This is probably due to the effects of foreign chromatin being highly nonuniform and influenced by diverse factors (Supplementary Note), in particular the wheat genetic background[27,28]. Taking advantage of the 'Lo7' chromosome-scale assembly, tracking of rye chromatin in wheat breeding programs will now become more reliable and predictable.

Rye vigor is in the genes. Rye distinguishes itself from other Triticeae through strong allogamy, which facilitates commercial hybrid-rye breeding, as well as conferring resilience to biotic stress and extreme winter-hardiness, qualifying rye as an important plant genetic resource in wheat improvement. Here, we showcase how the high-quality genome assembly sheds light on the genetic control of these specific aspects of rye biology.

Fertility restoration in rye and wheat. Rye hybrid breeding relies on efficient cytoplasmic male-sterility (CMS)/restorer-of-fertility (Rf) systems; however, the underlying molecular mechanisms have yet to be elucidated. Known *Rf* genes belong to a distinct clade of the pentatricopeptide repeat (PPR) RNA-binding factor family, referred to as *Rf*-like (RFL)[29,30]. Members of the mitochondrial transcription TERmination factor (mTERF) family are probably also involved in male fertility restoration in cereals[31,32]. The 'Lo7' assembly reveals a PPR-RFL/mTERF hotspot on 4RL coinciding with known *Rf* loci for two rye CMS systems known as CMS-P (the commercially predominant 'Pampa'-type) and CMS-C[12,33–35](Methods; Fig. 4a–f, Supplementary Tables 26 and 27 and Supplementary Note). We determined, as previously suggested, that these two loci, *Rfp* and *Rfc*, are closely linked but physically distinct[36] (Supplementary Table 28). Two members of the PPR-RFL clade reside within 0.186 Mb of the *Rfc1* locus (Supplementary Tables 26–28). The *Rfp* locus, in contrast, is neighbored by four *mTERF* genes (Supplementary Tables 27–28), in agreement with previous reports that an mTERF protein represents the *Rfp1* candidate gene in rye[32,37].

The new assembly also helped to dissect a strong candidate gene for the wheat locus Rf^{multi} (*Restoration-of-fertility in multiple CMS systems*) on wheat chromosome 1BS. Replacement of the wheat Rf^{multi} locus by its rye ortholog using 1RS.1BL chromosome translocations produces the male-sterile phenotype[38,39]. At the syntenic position of Rf^{multi}, wheat and rye share a PPR-RFL gene cluster[7] (Fig. 4k, Supplementary Table 26 and Supplementary Note). Only two wheat RFL-PPR genes in the cluster, *TraesCS1B02G071642.1* and *TraesCS1B02G072900.1*, encode full-length proteins; only the latter corresponds to a putative rye ortholog (*SECCE1Rv1G0008410.1*). Thus, the absence of a *TraesCS1B02G071642.1* ortholog in the nonrestorer rye suggest it as an attractive Rf^{multi} candidate. Current implementations of a wheat–rye Rf^{multi} CMS system involve 1RS.1BL translocations[5,40,41], which are typically linked to reduced baking quality[27]. Efforts to break this linkage may now benefit from marker development and/or genome editing approaches targeting *TraesCS1B02G071642.1* (ref. [42]).

Divergence of disease resistance loci in Triticeae. Rye plays an important role as genetic resource of biotic stress tolerance in wheat varieties carrying rye chromatin insertions. Race-specific pathogen resistance is typically associated with members of the class of nucleotide-binding-site and leucine-rich repeat (NLR)-motif genes[43]. We annotated 792 full-length rye NLR genes, finding them enriched in distal chromosomal regions, similar to what has been seen recently in the bread wheat genome[7,44] (Fig. 4e and Supplementary Tables 29 and 30). We compared the genomic regions in rye that are orthologous to mildew (*Pm2*, *Pm3* and *Mla*) and rust (*Lr10*) resistance gene loci from wheat and barley (Fig. 4g–j, Supplementary Table 31 and Supplementary Note). All loci, except for *Lr10*, contained complex gene families with several subfamilies either present or absent in

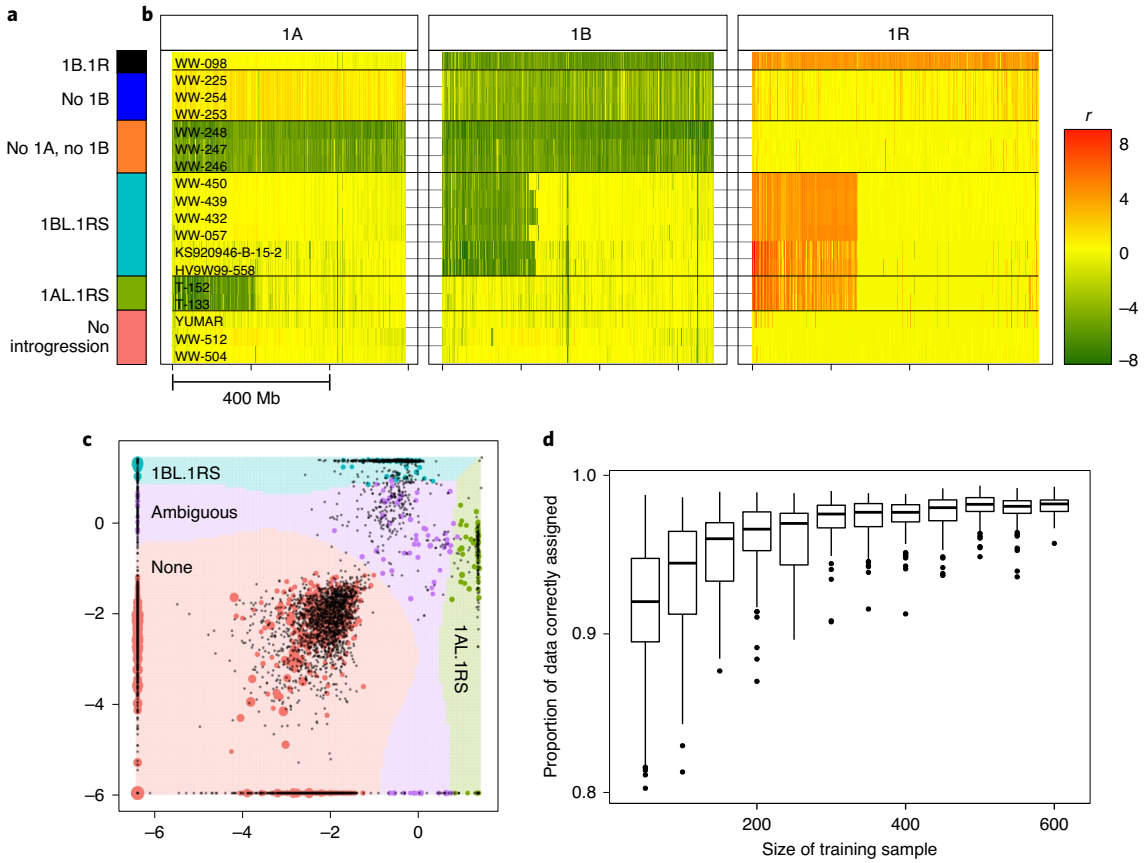

Fig. 3 | Combined reference mapping as a means to classify wheat and wheat–rye introgression karyotypes. a, Color key for **b** and **c**. **b**, Normalized read mapping depths for 1-Mb bins of chromosomes 1A, 1B and 1R, for a selection of wheat lines (including also some *Aegilops tauschii* accessions which contain no A or B subgenome) with various chromosome complements and introgressions (rows). The value *r* denotes the difference between the $\log_2$ reads per million mapping to a bin, compared to *T. aestivum* cv. 'Chinese Spring'. **c**, Visual representation of the SVM classifier, with the two selection features based on relative read mapping densities to 'translocation-prone' and 'other' chromosomal regions (Methods) shown on the *x* and *y* axes. Points represent training samples, with color corresponding to human-designated classification and size proportional to the total number of mapped reads for the sample. Black points are samples not classified by a human. Background colors represent the hypothetical classification that would be given to a sample at that position. **d**, Results of cross-validation testing the accuracy of the classifier and its relationship to the size of the training set. Box edges and whiskers represent quartiles and the center lines show the arithmetic means.

Table 2 | Summary of fixed effects estimates from linear mixed model estimating the influence of rye–wheat translocations upon yield, in two wheat diversity panels

Panel	Introgression type	Estimated yield effect	s.e.	*t*	Degrees of freedom	*P*
KSU	1AL.1RS	4.06%	0.54	7.54	1.89×10^4	4.95×10^{-14}
KSU	1BL.1RS	1.50%	0.41	3.68	1.88×10^4	0.00023
USDA	1AL.1RS	0.86%	0.31	2.72	2.82×10^4	0.0064
USDA	1BL.1RS	4.55%	0.39	11.78	2.82×10^4	$<2.0\times10^{-16}$

P values are calculated using a one-sided Student's *t*-test on the null hypothesis that the true yield effect is zero (Methods; Supplementary Table 25).

individual genomes, indicating either functional redundancy or the evolution of distinct resistance specificities or targets. For example, the wheat *Pm3* and rye *Pm8*/*Pm17* genes are orthologs and belong to a subfamily (clade A, Fig. 4i) which is absent in barley, whereas a different distinct subfamily (clade B, Fig. 4i) of the *Pm3* genes is present in wheat and barley but absent in rye (Supplementary Note). In essence, the 'Lo7' assembly reveals the genomic organization of conserved or nonorthologous NLR gene clusters, which can be exploited in future rye and wheat improvement efforts.

Frost tolerance. Rye possesses superior low-temperature tolerance (LTT) to other Triticeae crops[45]. A syntenic locus *Fr2* comprising a cluster of CBF (C-repeat/DRE-binding factor) genes is present on Triticeae group 5 chromosomes controlling LTT[46] in rye[47], *T. monococcum*[48], bread wheat[49,50] and barley[51]. In cold-tolerant varieties, LTT-implicated CBF genes of the *Fr2* locus are transcriptionally upregulated[52,53]. In 'Lo7', the *Fr2* locus comprises a cluster of 21 CBF-related genes at 614.3–616.5 Mb on 5R (Fig. 4f and Supplementary Table 32).

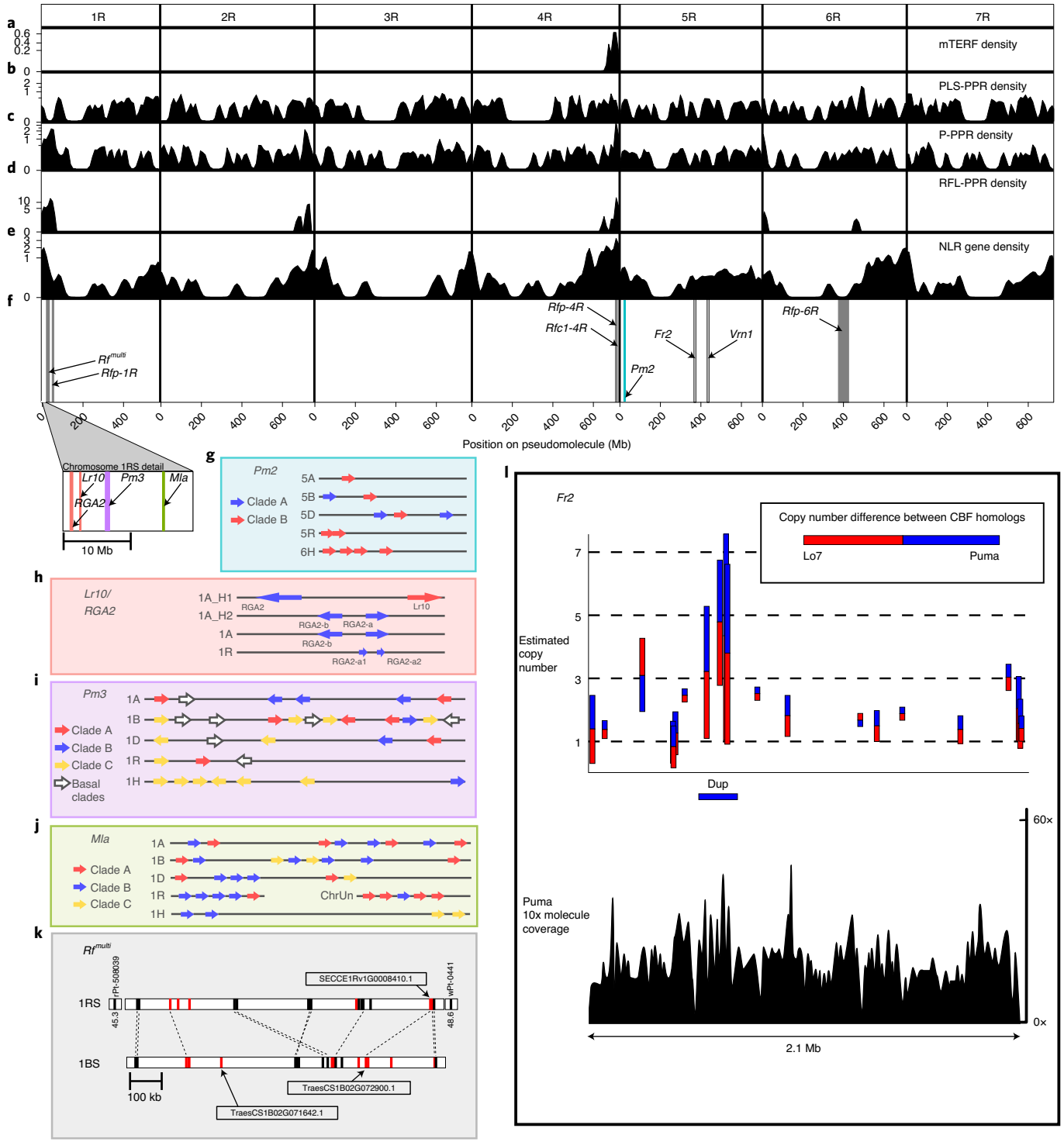

Fig. 4 | Comparative genomics of rye genes with agricultural importance. a–e, Density (instances per Mb) of mTERFs (**a**), PPRs (**d**) as well as NLRs (**e**) across the pseudomolecules. For visualization, the *y* axis is transformed using $x \rightarrow x^{1/3}$. **f**, Genomic locations of genes and loci discussed in the text. Colored bars correspond and refer to the colors of the box outlines in **g**–**k**; **g**–**j**, physical organization of selected NLR gene clusters compared across cultivated Triticeae genomes: *Pm2* (**g**), *Lr10/RGA2* (**h**), *Pm3* (**i**) and *Mla* (**j**). **k**, Organization of RFL genes at the 'Lo7' Rf^{multi} locus compared to its wheat ('Chinese Spring') counterpart. Flanking markers are shown on either end of the rye sequence. Two full-length wheat RFLs and a putative rye ortholog are labeled. PPR genes are colored red. **l**, CNV between 'PUMA-SK' and 'Lo7' within the *Fr2* interval revealed by 10x Genomics linked read sequencing. A (Dup)lication flagged by the Loupe analysis software is marked. The estimated copy number differences between 'Lo7' and 'Puma' are shown for *CBF* genes.

Since CBF gene family expansion correlates with increased LTT in other Triticeae[54] (Supplementary Note), we analysed phased-linked-read (10x Genomics Chromium) data of an *Fr2*-homozygous line with exceptional LTT ('Puma-SK' derived from rye variety 'Puma') in comparison to 'Lo7' (low LTT). Four of the *Fr2* CBF genes, all members of the same CBF subfamily (Supplementary Note) for which CNV has been previously implicated in LTT in wheat[54], showed patterns of copy number

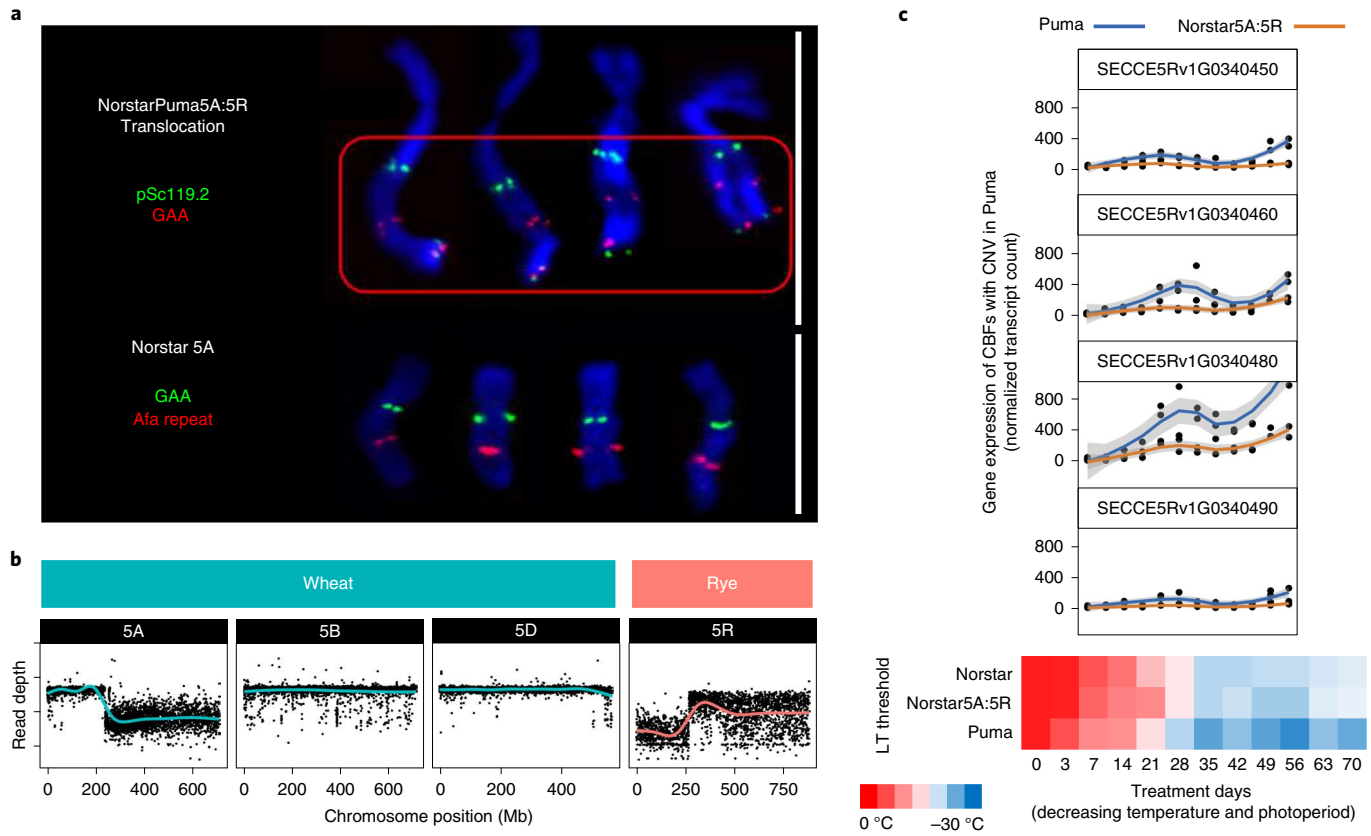

Fig. 5 | The cold tolerance associated region *Fr2* in 'Puma' and 'NorstarPuma5A:5R' translocation line. a, Chromosome labeling (top) using probes specific for 'Norstar' chromosome 5A (Afa) and 'Puma' 5R (pSc119.2) confirm the presence of a rye translocation in NorstarPuma5A:5R (red box), which also alters the binding of GAA. White bars, 10 µm. **b**, Combined reference read mapping of group 5 chromosomes confirms the balanced translocation event, gain of a large region of chromosome 5R from 'Puma' (rye, light read line) and loss of a large region on chromosome 5A of 'Norstar' (wheat, light blue line) in 'NorstarPuma5A:5R'. Read depth is given in $\log_2$ reads per million versus 'Chinese Spring'. **c**, Gene expression analysis of rye CBF genes with CNV in 'Puma' (blue line) and 'NorstarPuma5A:5R' (orange line). Plants were grown in a time series with decreasing day length and temperature over a 70-d period and the temperatures at which 50% lethality was observed (LT50) were recorded (heatmap).

variation (CNV) (*SECCE5Rv1G030450*, *SECCE5Rv1G030460*, *SECCE5Rv1G030480* and *SECCE5Rv1G030490*; Supplementary Table 33 and Supplementary Note).

Transferring superior LTT from rye to wheat by translocation is an attractive breeding goal. We derived a 5A.5RL translocation line in winter wheat 'Norstar' using 'Puma' rye as the 5R donor, thus replacing the wheat 5A CBF cluster (Methods; Fig. 5a,b). LTT, however, was not notably altered by the translocation compared to 'Norstar' (Fig. 5c), suggesting that the rye CBF gene cluster is activated but, as previously suggested by Campoli et al.[53], differently regulated in the wheat background. Gene expression of 'Puma' CBFs with CNV were indeed attenuated during treatments of cold stress in 'Norstar5A:5R' (Methods; Fig. 5c). Therefore, transferring LTT from rye into wheat will require indepth understanding of differences in the LTT regulatory network between rye and wheat.

Discussion

The high-quality chromosome-scale assembly of rye inbred line 'Lo7' constitutes an important step forward in genome analysis of the Triticeae crop species and complements the resources recently made available for different wheat species[14,55–58] and barley[13,59]. This resource will help reveal the genomic basis of differences in major life-history traits between the self-incompatible, cross-pollinating rye and its selfing and inbreeding relatives. Our evolutionary analyses demonstrate that rye subspecies are better conceptualized as a reticulated group of incipient species and that mechanisms such as transposable-element expansion and SV between genotypes are probably acting to bring about evolutionary divergence. The joint use of the rye and wheat genomes to characterize the effects of rye chromatin introgressions may provide a short-term opportunity to breeders as they continue to better separate confounding variables from the genetic combinations that best improve yield in various environments; but these benefits will ultimately be affected by negative linkage so long as whole chromosome arm translocations are involved. Discoveries at the single-gene level—such as the contributions offered here to pathogen resistance, LTT and male fertility restoration control—will be best tested and exploited by finer-scale manipulation in dedicated experiments[12]. This is an indispensable prerequisite for the development of gene-based strategies that exploit untapped genetic diversity in breeding materials and ex situ gene banks to improve small grain cereals and meet the changing demands of global environments, farmers and society.

Online content

Any methods, additional references, Nature Research reporting summaries, source data, extended data, supplementary information, acknowledgements, peer review information; details of author contributions and competing interests; and statements of data and code availability are available at https://doi.org/10.1038/s41588-021-00807-0.

Received: 6 December 2019; Accepted: 29 January 2021;
Published online: 18 March 2021

References

1. Beck, H. E. et al. Present and future Köppen–Geiger climate classification maps at 1-km resolution. *Sci. Data* **5**, 180214 (2018).
2. Sharma, S. et al. Integrated genetic map and genetic analysis of a region associated with root traits on the short arm of rye chromosome 1 in bread wheat. *Theor. Appl. Genet.* **119**, 783–793 (2009).
3. Lukaszewski, A.J. in *Alien Introgression in Wheat* (eds Molnár-Láng, M. et al.) 163–189 (Springer, 2015).
4. Kim, W., Johnson, J., Baenziger, P., Lukaszewski, A. & Gaines, C. Agronomic effect of wheat–rye translocation carrying rye chromatin (1R) from different sources. *Crop Sci.* **44**, 1254–1258 (2004).
5. Crespo-Herrera, L. A., Garkava-Gustavsson, L. & Åhman, I. A systematic review of rye (*Secale cereale* L.) as a source of resistance to pathogens and pests in wheat (*Triticum aestivum* L.). *Hereditas* **154**, 14 (2017).
6. Doležel, J. et al. Plant genome size estimation by flow cytometry: inter-laboratory comparison. *Ann. Bot.* **82**, 17–26 (1998).
7. IWGSC. Shifting the limits in wheat research and breeding using a fully annotated reference genome. *Science* **361**, eaar7191 (2018).
8. Martis, M. M. et al. Reticulate evolution of the rye genome. *Plant Cell* **25**, 3685–3698 (2013).
9. Bauer, E. et al. Towards a whole-genome sequence for rye (*Secale cereale* L.). *Plant J.* **89**, 853–869 (2017).
10. Schneider, A., Rakszegi, M., Molnár-Láng, M. & Szakács, É. Production and cytomolecular identification of new wheat-perennial rye (*Secale cereanum*) disomic addition lines with yellow rust resistance (6R) and increased arabinoxylan and protein content (1R, 4R, 6R). *Theor. Appl. Genet.* **129**, 1045–1059 (2016).
11. Li, J., Zhou, R., Endo, T. R. & Stein, N. High-throughput development of SSR marker candidates and their chromosomal assignment in rye (*Secale cereale* L.). *Plant Breed.* **137**, 561–572 (2018).
12. Hackauf, B. et al. QTL mapping and comparative genome analysis of agronomic traits including grain yield in winter rye. *Theor. Appl. Genet.* **130**, 1801–1817 (2017).
13. Mascher, M. et al. A chromosome conformation capture ordered sequence of the barley genome. *Nature* **544**, 427–433 (2017).
14. Maccaferri, M. et al. Durum wheat genome highlights past domestication signatures and future improvement targets. *Nat. Genet.* **51**, 885 (2019).
15. Zhu, T. et al. Improved genome sequence of wild emmer wheat Zavitan with the aid of optical maps. *G3* **9**, 619–624 (2019).
16. Zimin, A. V. et al. Hybrid assembly of the large and highly repetitive genome of *Aegilops tauschii*, a progenitor of bread wheat, with the MaSuRCA mega-reads algorithm. *Genome Res.* **27**, 787–792 (2017).
17. Braun, E.-M. et al. Gene expression profiling and fine mapping identifies a gibberellin 2-oxidase gene co-segregating with the dominant dwarfing gene *Ddw1* rye (*Secale cereale* L.). *Front. Plant Sci.* **10**, 857 (2019).
18. Lieberman-Aiden, E. et al. Comprehensive mapping of long-range interactions reveals folding principles of the human genome. *Science* **326**, 289–293 (2009).
19. Wicker, T. et al. A unified classification system for eukaryotic transposable elements. *Nat. Rev. Genet.* **8**, 973–982 (2007).
20. Ou, S., Chen, J. & Jiang, N. Assessing genome assembly quality using the LTR Assembly Index (LAI). *Nucleic Acids Res.* **46**, e126 (2018).
21. Wicker, T., Gundlach, H. & Schulman, A.H. in *The Barley Genome* (eds SteinGary, N. & J. Muehlbauer, J.) 123–138 (Springer, 2018).
22. Wicker, T. et al. Impact of transposable elements on genome structure and evolution in bread wheat. *Genome Biol.* **19**, 103 (2018).
23. Himmelbach, A. et al. Discovery of multi-megabase polymorphic inversions by chromosome conformation capture sequencing in large-genome plant species. *Plant J.* **96**, 1309–1316 (2018).
24. Lowry, D.B. & Willis, J.H. A widespread chromosomal inversion polymorphism contributes to a major life-history transition, local adaptation, and reproductive isolation. *PLoS Biol.* **8**, e1000500 (2010).
25. Schreiber, M., Himmelbach, A., Börner, A. & Mascher, M. Genetic diversity and relationship between domesticated rye and its wild relatives as revealed through genotyping-by-sequencing. *Evol. Appl.* **12**, 66–77 (2019).
26. Friebe, B., Jiang, J., Raupp, W., McIntosh, R. & Gill, B. Characterization of wheat-alien translocations conferring resistance to diseases and pests: current status. *Euphytica* **91**, 59–87 (1996).
27. Graybosch, R. A. Mini review: uneasy unions: quality effects of rye chromatin transfers to wheat. *J. Cereal Sci.* **33**, 3–16 (2001).
28. Kumlay, A. et al. Understanding the effect of rye chromatin in bread wheat. *Crop Sci.* **43**, 1643–1651 (2003).
29. Chen, L. & Liu, Y.-G. Male sterility and fertility restoration in crops. *Annu. Rev. Plant Biol.* **65**, 579–606 (2014).
30. Melonek, J., Stone, J. D. & Small, I. Evolutionary plasticity of restorer-of-fertility-like proteins in rice. *Sci. Rep.* **6**, 35152 (2016).
31. Bernhard, T., Koch, M., Snowdon, R. J., Friedt, W. & Wittkop, B. Undesired fertility restoration in *msm1* barley associates with two mTERF genes. *Theor. Appl. Genet.* **132**, 1335–1350 (2019).
32. Hackauf, B., Korzun, V., Wortmann, H., Wilde, P. & Wehling, P. Development of conserved ortholog set markers linked to the restorer gene Rfp1 in rye. *Mol. Breed.* **30**, 1507–1518 (2012).
33. Geiger, H., Yuan, Y., Miedaner, T. & Wilde, P. Environmental sensitivity of cytoplasmic genic male sterility (CMS) in *Secale cereale* L. *Fortschr. Pflanzen.* **18**, 7–18 (1995).
34. Geiger, H. Cytoplasmatisch-genische pollensterilität in roggenformen iranischer abstammung. *Naturwissenschaften* **58**, 98–99 (1971).
35. Geiger, H. & Schnell, F. Cytoplasmic male sterility in rye (*Secale cereale* L.) 1. *Crop Sci.* **10**, 590–593 (1970).
36. Stojałowski, S., Jaciubek, Moa & Masojć, P. Rye SCAR markers for male fertility restoration in the P cytoplasm are also applicable to marker-assisted selection in the C cytoplasm. *J. Appl. Genet.* **46**, 371–373 (2005).
37. Wilde, P. et al. Restorer plants. US patent application 16/064,304 (2019).
38. Tsunewaki, K. Fine mapping of the first multi-fertility-restoring gene, Rf^{multi}, of wheat for three *Aegilops* plasmons, using 1BS-1RS recombinant lines. *Theor. Appl. Genet.* **128**, 723–732 (2015).
39. Hohn, C. E. & Lukaszewski, A. J. Engineering the 1BS chromosome arm in wheat to remove the *Rf* multi locus restoring male fertility in cytoplasms of *Aegilops kotschyi*, *Ae. uniaristata* and *Ae. mutica*. *Theor. Appl. Genet.* **129**, 1769–1774 (2016).
40. Jung, W. J. & Seo, Y. W. Employment of wheat–rye translocation in wheat improvement and broadening its genetic basis. *J. Crop Sci. Biotechnol.* **17**, 305–313 (2014).
41. Lukaszewski, A. J. Chromosomes 1BS and 1RS for control of male fertility in wheats and triticales with cytoplasms of *Aegilops kotschyi*, *Ae. mutica* and *Ae. uniaristata*. *Theor. Appl Genet* **130**, 2521–2526 (2017).
42. Hensel, G. Genetic transformation of Triticeae cereals—summary of almost three-decade's development. *Biotechnol. Adv.* **40**, 107484 (2020).
43. Kourelis, J. & van der Hoorn, R. A. Defended to the nines: 25 years of resistance gene cloning identifies nine mechanisms for R protein function. *Plant Cell* **30**, 285–299 (2018).
44. Steuernagel, B. et al. The NLR-Annotator tool enables annotation of the intracellular immune receptor repertoire. *Plant Physiol.* **183**, 468–482 (2020).
45. Dvorak, J. & Fowler, D. Cold hardiness potential of triticale and teraploid Rye 1. *Crop Sci.* **18**, 477–478 (1978).
46. Jung, W. J. & Seo, Y. W. Identification of novel C-repeat binding factor (CBF) genes in rye (*Secale cereale* L.) and expression studies. *Gene* **684**, 82–94 (2019).
47. Börner, A., Korzun, V., Voylokov, A., Worland, A. & Weber, W. Genetic mapping of quantitative trait loci in rye (*Secale cereale* L.). *Euphytica* **116**, 203–209 (2000).
48. Vágújfalvi, A., Galiba, G., Cattivelli, L. & Dubcovsky, J. The cold-regulated transcriptional activator *Cbf3* is linked to the frost-tolerance locus *Fr-A2* on wheat chromosome 5A. *Mol. Genet. Genomics* **269**, 60–67 (2003).
49. Båga, M. et al. Identification of quantitative trait loci and associated candidate genes for low-temperature tolerance in cold-hardy winter wheat. *Funct. Integr. Genom.* **7**, 53–68 (2007).
50. Fowler, D., N'Diaye, A., Laudencia-Chingcuanco, D. & Pozniak, C. Quantitative trait loci associated with phenological development, low-temperature tolerance, grain quality, and agronomic characters in wheat (*Triticum aestivum* L.). *PLoS ONE* **11**, e0152185 (2016).
51. Francia, E. et al. Two loci on chromosome 5H determine low-temperature tolerance in a 'Nure'(winter)×'Tremois'(spring) barley map. *Theor. Appl. Genet.* **108**, 670–680 (2004).
52. Stockinger, E. J., Gilmour, S. J. & Thomashow, M. F. *Arabidopsis thaliana* CBF1 encodes an AP2 domain-containing transcriptional activator that binds to the C-repeat/DRE, a cis-acting DNA regulatory element that stimulates transcription in response to low temperature and water deficit. *Proc. Natl Acad. Sci. USA* **94**, 1035–1040 (1997).
53. Campoli, C., Matus-Cádiz, M. A., Pozniak, C. J., Cattivelli, L. & Fowler, D. B. Comparative expression of *Cbf* genes in the Triticeae under different acclimation induction temperatures. *Mol. Genet. Genomics* **282**, 141–152 (2009).
54. Würschum, T., Longin, C. F. H., Hahn, V., Tucker, M. R. & Leiser, W. L. Copy number variations of CBF genes at the *Fr-A2* locus are essential components of winter hardiness in wheat. *Plant J.* **89**, 764–773 (2017).
55. Avni, R. et al. Wild emmer genome architecture and diversity elucidate wheat evolution and domestication. *Science* **357**, 93–97 (2017).
56. Ling, H.-Q. et al. Draft genome of the wheat A-genome progenitor *Triticum urartu*. *Nature* **496**, 87–90 (2013).
57. Luo, M. et al. Genome sequence of the progenitor of the wheat D genome *Aegilops tauschii*. *Nature* **551**, 498–502 (2017).
58. IWGSC. A chromosome-based draft sequence of the hexaploid bread wheat (*Triticum aestivum*) genome. *Science* **345**, 1251788 (2014).
59. Monat, C. et al. TRITEX: chromosome-scale sequence assembly of Triticeae genomes with open-source tools. *Genome Biol.* **20**, 284 (2019).

Methods

Genome size estimation by flow cytometry. We characterized the landscape of *S. cereale* genome sizes to contextualize the size of the 'Lo7' genome and gain an impression of genome size variation within the species. Grains from 15 diverse rye accessions were from nine providers listed in Supplementary Table 1. Plants of pea (seeds provided by Semo breeding station, Smržice, Czech Republic) served as an internal reference standard in flow cytometric estimation of nuclear DNA content in all accessions, except of the tetraploid accession ACE-1, for which *S. cereale* line 'Lo7' was used as a reference. Plants were raised in garden compost in pots and maintained in a greenhouse until they reached a height of 10–20 cm. Nuclear genome size was estimated essentially as described by Doležel et al.[60] using a CyFlow Space flow cytometer (Sysmex Partec) equipped with a 532-nm green laser. The gain of the instrument was adjusted so that the peak representing interphase first growth (G1) nuclei of the standard was positioned approximately on channel 100 on a histogram of relative fluorescence intensity when using a 512-channel scale. Five individual plants per test species were sampled and each sample was analysed three times, each time on a different day. A minimum of 5,000 nuclei per sample was analysed and 2C DNA contents in pg were calculated from the means of the G1 peak positions by applying the following formula:

$$\text{2C nuclear DNA content} = \frac{\text{sample G1 peak mean} \times \text{standard 2C DNA content}}{\text{standard G1 peak mean}}$$

Mean nuclear DNA content (2C) was then calculated for each accession. DNA contents in pg were converted to genome size in bp using the conversion factor 1 pg DNA = 0.978 Gb (ref. [61]). Statistical analysis was performed using NCSS 97 statistical software package (Statistical Solutions). One-way analysis of variance and a Bonferroni's (all pairwise) multiple comparison test were used for analysis of variation in monoploid (1C×) genome size. A significance level $\alpha = 0.01$ was used.

'Lo7' genome assembly and gene annotation. Descriptions of the assembly methods, descriptions of the data generation and the annotation procedure for gene features, are given in the Supplementary Note.

Fluorescence in situ hybridization (FISH). Three-day-old roots of the rye accession were pretreated in 0.002 M 8-hydroxyquinoline at 7 °C for 24 h and fixed in ethanol:acetic acid (3:1 v/v). Chromosome preparation and FISH were performed according to the methods described by Aliyeva-Schnorr et al.[62]. The hybridization mixture contained 50% deionized formamide, 2× SSC, 20% dextran sulfate and 5 ng µl^{-1} of each probe. Slides were denatured at 75 °C for 3 min and the final stringency of hybridization was 76%. We used 34–65 nucleotide-long 5′-labeled oligo probes designed for the in silico identified repeats and published probe sequence (Supplementary Table 21). Images were captured using an epifluorescence microscope BX61 Olympus equipped with a cooled CCD camera (Orca-ER, Hamamatsu). Chromosomes were identified visually on the basis primarily of morphology, heterochromatic DAPI+ bands and the localization of probe pSc119.2.1 (ref. [63]) (Supplementary Note).

Gene-level synteny and percentage identity scores between rye and other Triticeae species. HC gene sequences from the 'Lo7' gene annotation were aligned to the annotated transcriptomes of bread wheat[7] (*T. aestivum* cv. 'Chinese Spring') and barley[13] (*H. vulgare* cv. 'Morex') using BLASTn (v.2.9.0+)[64] with default parameters. The lowest E-value alignment for each gene against the transcriptome associated with each subject genome (or subgenome) was selected, with the highest bitscore and then longest alignment chosen in the case of a tie. Only reciprocal best matches per (sub)genome were accepted. Relative evolutionary distances between rye, barley and the wheat subgenomes were estimated using the mean percentage identity scores of these filtered matches, calculated in bins of 100 reciprocal matches (in increments of 20 bins). The positions of the bins on the pseudomolecules were taken to be the mean match position of the matches within each bin.

Phylogenetic analyses, IBS statistics, F_{st}, *D*-statistics and P_n/P_s. The genotyping-by-sequencing (GBS) dataset of 603 samples from Schreiber et al.[25] was extended by a 347 further GBS samples from the IPK gene bank (mainly wild *Secale* taxa) and the five samples used in the Hi-C SV-detection study ('Lo7', 'Lo225', 'R1003', 'R925' and 'R2446'). The resulting sample set ($n = 955$) and passport data are listed in Supplementary Table 34. DNA isolated from the five Hi-C samples was sent to Novogene (en.novogene.com/) for Illumina library construction and sequencing in multiplex on the NovaSeq platform (paired-end 150-bp reads, ~140 Gb per sample, S2 flow cell). Demultiplexing, adapter trimming, read mapping and variant calling correspond to the approach described in Schreiber et al.[25], using the new reference for read mapping. The dataset was filtered for a maximum of 30% missing data and a minor allele frequency of 1% resulting in 72,465 SNPs was used (Supplementary Note). A neighbor joining tree was constructed with the R package 'ape'[65], on the basis of genetic distances computed with the R package SNPRelate[66]. Principal component analysis (PCA) was performed with smartPCA from the EIGENSOFT v.6.0.1 package (github.com/DReichLab/EIG) using least square projection without outlier removal. Seven rye genetic clusters were designated using the 'kmeans' R function, with default parameters, using the first three principal components from the PCA as input.

For IBS and F_{st} analyses, a more stringent filtering regime requiring read depth ≥ 6, maximum 5% missing data and call quality ≥ 250 (resulting in 9,538 SNPs) was selected (Supplementary Note). IBS scores between 'Lo7' and the other lines in the set were calculated in windows of 100 consecutive SNP loci (at intervals of 25 SNP loci) using the snpgdsIBS function in the R package SNPRelate. F_{st} was calculated in the same windows using the snpgdsFst function in the R package SNPRelate (using method='W&H02'). The calculation was performed for every pairwise combination of the following groups: 'Domesticated-like' (cluster 3), 'Wild–*S. strictum*-like' (clusters 1, 5 and 6) and 'Wild–*S. sylvestre*-like' (cluster 4).

To assign ancestral states to variant SNPs segregating in rye, exonic variants identified in the GBS dataset were coupled to their orthologs alleles in the three bread wheat ('Chinese Spring')[7] alleles using the rye-versus-wheat CDS BLAST alignments (see above) and parsing the BLAST alignment strings using the custom script blast_get_alleles_at_position.c (https://github.com/mtrw/tim_bioinfo_tools). Reciprocal best matches were calculated separately for the alignments between the rye CDS set and the set of CDSs from each wheat genome. The ancestral allele was assigned by consensus among the wheat genomes and, if no allele claimed a majority, the variant was omitted from the dataset. Genome-wide *D*-statistics were calculated according to the four-taxon ABBA-BABA method as described in ref. [67], with the wheat consensus allele as the outgroup and selections of the *k*-means-assigned clusters selected as the three test populations. Estimator variance was approximated via the block jackknife procedure, with 5 Mb exclusion bins. The effects of the rye-versus-wheat nucleotide differences falling within coding sequences were annotated using SnpEff (v.4, 'ann' function), with default parameters. P_n/P_s scores (the ratio of counts of nonsynonymous to synonymous differences to wheat in variants segregating in rye) were calculated using the same binning scheme as was used for F_{st} and IBS (see above). P_n/P_s scores were calculated separately for subgroups of rye clusters representing 'wild-like' and 'domesticated-like' ryes separately (see main text). P_n/P_s scores were only estimated for bins in which the combined number of rye-segregating variants exceeded nine.

Wheat–rye introgression haplotype identification and classification. We assayed for the presence of 1R germplasm in wheat genotypes in silico by mapping various wheat sequence data to a combined reference genome made up of the pseudomolecules of rye line 'Lo7' (this study) and wheat cv. 'Chinese Spring'[7]. Publicly available data were obtained from WHEALBI project resources[68] ($n = 506$), CIMMYT ($n = 903$) and KSU ($n = 4{,}277$). GBS libraries were constructed and sequenced for samples from USDA-RPN ($n = 875$; Supplementary Table 23) as described in Rife et al.[69]. On the basis of the approach described by Keilwagen et al.[70], reads were demultiplexed with a custom C script (github.com/umngao/splitgbs) and aligned to the combined reference using bwa mem (v.0.7, arguments -M)[71] after trimming adapters with cutadapt[72]. The aligned reads from all panels were filtered for quality using samtools[73] (v.1.9, arguments -F3332 -q20). The numbers of reads aligned to 1 Mb nonoverlapping bins on each pseudomolecule were tabulated. The counts were expressed as rpmm ≡ $\log_2$(reads mapped to bin per million reads mapped). To control for mappability biases over the genome, the rpmm for each bin was normalized by subtracting the rpmm attained by the 'Chinese Spring' sample for the same bin to give the normalized rpmm, *r*.

To investigate the possibility of classifying the samples automatically, visual representations of *r* across the combined reference genome were inspected and obvious cases of 1R.1A and 1R.1B introgression were distinguished from several other karyotypes, including nonintrogressed samples and ambiguous samples showing a slight overabundance of 1RS reads but less discernible signals of depletion in 1A or 1B (Supplementary Note). We defined the following features for each sample: $featureA = -\log[(\text{mean}(r^{1A}_I) - \text{mean}(r^{1A}_N)) \times (\text{mean}(r^{1R}_I) - \text{mean}(r^{1R}_N))]$ and $featureB = -\log[(\text{mean}(r^{1B}_I) - \text{mean}(r^{1B}_N)) \times (\text{mean}(r^{1R}_I) - \text{mean}(r^{1R}_N))]$. Whenever the term inside the log was negative (and would thus give an undefined result), the value of the feature was set to the minimum of the defined values for that feature. The quantity $\text{mean}(r^{1R}_I)$ refers to the average value of *r* for all bins within the terminal 200 Mb of the normally introgressed (*I*) end of 1R (an *N* in the subscript denotes the terminal 300 Mb of the normally nonintrogressed (*N*) arm) and so forth for other chromosomes. This choice of feature definition meant that, wherever little difference in *r* occurred between 1RS and 1RL, suggesting no presence of rye, the factor $\text{mean}(r^{1R}_I) - \text{mean}(r^{1R}_N)$ would pull the feature values close to the origin and differences between *r* on the long and short arms of 1A or 1B would pull the values of A or B respectively away from the origin, depending upon which introgressions are present. A classifier was developed by training an SVM to distinguish nonintrogressed, 1A.1R-introgressed, 1B.1R-introgressed and ambiguously introgressed samples, using the function ksvm (arguments Type='C-svc', kernel='rbfdot', C=1) from the R package kernlab. Classification results are given in Supplementary Table 25. Testing was performed by generating sets of between 50 and 600 random samples from the dataset and using these to train a model, then using the kernlab::predict() to test the model's accuracy of prediction on the remaining data not used in training. This was repeated 100 times for each training dataset size.

To confirm the common origin of the 1AL.RS and 1BL.1RS introgressions, predicted 1RS carriers were selected to form a combined 1RS panel (over 1,200

lines) to call SNPs. A total of over 3 million SNPs were called with samtools/bcftools v.1.9 (mpileup -q20, -r chr1R:1-300000000; call -mv). SNPs were filtered on the basis of combined minimum read depth of 25, minor allele frequency of 0.01. A total of >900,000 SNPs were obtained. All IBS percentages were calculated and the square root values of per cent different calls were used to derive a heatmap for all pairwise comparisons (Supplementary Note).

SV detection in 'Puma-SK' and 'NorstarPuma5A:5R'. To characterize the *Fr2* region in 'Puma-SK' and the introgression in 'NorstarPuma5A:5R', whole-genome sequencing was performed using the Chromium 10x Genomics platform. Nuclei were isolated from 30 seedlings and high molecular-weight genomic DNA was extracted from nuclei using phenol chloroform according to the protocol of Zheng et al.[74]. Genomic DNA was quantified by fluorometry using Qubit 2.0 Broad Range (Thermofisher) and size selection was performed to remove fragments smaller than 40 kb using pulsed field electrophoresis on a Blue Pippin (Sage Science) according to the manufacturer's specifications. Integrity and size of the size-selected DNA were determined using a Tapestation 2200 (Agilent) and Qubit 2.0 Broad Range (Thermofisher), respectively. Library preparation was performed as per the 10x Genomics Genome Library protocol (https://support.10xgenomics.com/genome-exome/library-prep/doc/user-guide-chromium-genome-reagent-kit-v2-chemistry) and uniquely barcoded libraries were prepared and multiplexed for sequencing by Illumina HiSeq. Demultiplexing and the generation of fastq files were performed using LongRanger v.2.2.0 mkfastq (https://support.10xgenomics.com/genome-exome/software/pipelines/latest/using/mkfastq; default parameters).

Sequencing reads from 'Puma-SK' and 'NorstarPuma5A:5R' were aligned to the rye line 'Lo7' and bread wheat cv. 'Chinese Spring'[7] genome assemblies, respectively, using LongRanger WGS (https://support.10xgenomics.com/genome-exome/software/pipelines/latest/using/wgs; arguments -vcmode 'freebayes'). Large-scale structural variants detected by LongRanger were visualized with a combination of Loupe (v.2.1.1; https://support.10xgenomics.com/genome-exome/software/visualization/latest/what-is-loupe; downloaded February 2019; Supplementary Table 33). Short variants were called using the Freebayes software (github.com/ekg/freebayes) implemented within the Longranger v.2.2.0 WGS pipeline. For determining the introgression, 'NorstarPuma5A5R' reads which did not map to the 'Chinese Spring' reference were aligned to the 'Lo7' assembly using the LongRanger align pipeline (https://support.10xgenomics.com/genome-exome/software/pipelines/latest/advanced/other-pipelines). Samtools (v.1.9)[73] bedcov was used to calculate the genome-wide read coverage across both references. CNV between 'Puma-SK' and 'Lo7' was detected using a combination of barcode coverage analysis output by the Longranger WGS pipeline and read depth-of-coverage based analysis using CNVnator[75] v.0.4 and cn.mops[76] v.1.12.0.

Expression profiling of 'NorstarPuma5A:5R' and 'Puma'. RNA from 'NorstarPuma5A:5R' and 'Puma' was isolated and sequenced as described above. Sequencing adapters were removed and low-quality reads were trimmed using Trimmomatic[77]. RNA reads from 'NorstarPuma5A:5R' and 'Puma' were aligned to the 'Lo7' reference using Hisat2 (ref. [78]; v.2.1.0; default arguments) and transcripts were quantified with htseq (ref. [79]; v.11.1; default parameters). Differential expression analysis was carried out using DESeq2 (ref. [80]; v.3.11; default parameters).

Reporting Summary. Further information on research design is available in the Nature Research Reporting Summary linked to this article.

Data availability

The 'Lo7' assembly and gene feature annotation data are available via e!DAL at https://doi.org/10.5447/ipk/2020/33 and https://doi.org/10.5447/ipk/2020/29. The visual suite of resources for assembly assessment are stored at https://doi.org/10.5447/ipk/2020/32. Raw sequence data generated in the course of the study are available at European Nucleotide Archive (ENA) with accession numbers PRJEB32636 (PE and MP data for assembly), PRJEB32574 and PRJEB34626 (Hi-C), PRJEB34439 (10x), PRJEB32587 (CSS), PRJEB35392 (GBS data) and PRJEB35461 (RNAseq and IsoSeq for annotation of 'Lo7'). Chromium 10x and RNAseq data for 'Puma' and 'Norstar' are available at PRJNA564622. The SNP matrix used for rye population genetic analyses is available via e!DAL at https://doi.org/10.5447/ipk/2020/31. GBS and sequence data generated for the USDA and CIMMYT wheat diversity panels are available at ENA with accession numbers PRJNA566410, PRJNA566408 and PRJNA566409. Optical map data and alignments are available via e!DAL at https://doi.org/10.5447/ipk/2020/30. High-stringency transposable element annotations (used for evolutionary analyses) are given in Supplementary Table 10, while the larger, low-stringency annotations (used for assembly quality comparisons) are available via e!DAL at https://doi.org/10.5447/ipk/2020/34.

Code availability

The custom miRNA manipulation scripts used in miRNA annotation (SumirFind.pl, SumirFold.pl, SumirLocate_v2.py and Sumircreen_v2.py) are available at https://github.com/hikmetbudak/miRNA-annotation. Two custom scripts used for parsing BLAST output (get_alleles_at_position.c and blast_to_snps.c) are available at https://github.com/mtrw/tim_bioinfo_tools and custom R functions extending or modifying functions of the TRITEX assembly pipeline (version corresponding to commit ID 2898e74) are available at https://github.com/mtrw/Sc_genome_assembly. The custom tool used to demultiplex wheat panel GBS data (splitgbs.c) is available at github.com/umngao/splitgbs.

References

60. Dolezel, J., Kubaláková, M., Paux, E., Bartos, J. & Feuillet, C. Chromosome-based genomics in the cereals. *Chromosome Res.* **15**, 51–66 (2007).
61. Dolezel, J., Bartos, J., Voglmayr, H. & Greilhuber, J. Nuclear DNA content and genome size of trout and human. *Cytom. A* **51**, 127–128 (2003).
62. Aliyeva-Schnorr, L., Ma, L. & Houben, A. A fast air-dry dropping chromosome preparation method suitable for FISH in plants. *J. Vis. Exp.* **16**, e53470 (2015).
63. Cuadrado, A., Jouve, N. & Ceoloni, C. Variation in highly repetitive DNA composition of heterochromatin in rye studied by fluorescence *in situ* hybridization. *Genome* **38**, 1061–1069 (1995).
64. Altschul, S. F., Gish, W., Miller, W., Myers, E. W. & Lipman, D. J. Basic local alignment search tool. *J. Mol. Biol.* **215**, 403–410 (1990).
65. Paradis, E. & Schliep, K. ape 5.0: an environment for modern phylogenetics and evolutionary analyses in R. *Bioinformatics* **35**, 526–528 (2018).
66. Zheng, X. et al. A high-performance computing toolset for relatedness and principal component analysis of SNP data. *Bioinformatics* **28**, 3326–3328 (2012).
67. Green, R. E. et al. A draft sequence of the Neandertal genome. *Science* **328**, 710–722 (2010).
68. Pont, C. et al. Tracing the ancestry of modern bread wheats. *Nat. Genet.* **51**, 905–911 (2019).
69. Rife, T.W., Graybosch, R.A. & Poland, J.A. Genomic analysis and prediction within a US public collaborative winter wheat regional testing nursery. *Plant Genome* **11**, e180004 (2018).
70. Keilwagen, J. et al. Detecting large chromosomal modifications using short read data from genotyping-by-sequencing. *Front. Plant Sci.* **10**, 1133 (2019).
71. Li, H. & Durbin, R. Fast and accurate long-read alignment with Burrows–Wheeler transform. *Bioinformatics* **26**, 589–595 (2010).
72. Martin, M. Cutadapt removes adapter sequences from high-throughput sequencing reads. *EMBnet J.* **17**, 10–12 (2011).
73. Li, H. et al. The sequence Alignment/Map format and SAMtools. *Bioinformatics* **25**, 2078–2079 (2009).
74. Zhang, M. et al. Preparation of megabase-sized DNA from a variety of organisms using the nuclei method for advanced genomics research. *Nat. Protoc.* **7**, 467–478 (2012).
75. Abyzov, A., Urban, A. E., Snyder, M. & Gerstein, M. CNVnator: an approach to discover, genotype, and characterize typical and atypical CNVs from family and population genome sequencing. *Genome Res.* **21**, 974–984 (2011).
76. Klambauer, G. et al. cn. MOPS: mixture of poissons for discovering copy number variations in next-generation sequencing data with a low false discovery rate. *Nucleic Acids Res.* **40**, e69 (2012).
77. Bolger, A. M., Lohse, M. & Usadel, B. Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics* **30**, 2114–2120 (2014).
78. Kim, D., Langmead, B. & Salzberg, S. L. HISAT: a fast spliced aligner with low memory requirements. *Nat. Methods* **12**, 357–360 (2015).
79. Anders, S., Pyl, P. T. & Huber, W. HTSeq—a Python framework to work with high-throughput sequencing data. *Bioinformatics* **31**, 166–169 (2015).
80. Love, M. I., Huber, W. & Anders, S. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biol.* **15**, 550 (2014).

Acknowledgements

We are grateful to M. Knauft, I. Walde, S. Koenig, M. Ziems and S. Thumm (Leibniz Institute of Plant Genetics and Crop Plant Research, IPK), J. Ens (University of Saskatchewan), C. Uauy and J. Simmonds (John Innes Centre), S. Duncan (Earlham Institute), Z. Dubská and J. Weiserová (Institute of Experimental Botany), A. Hastie (Bionano Genomics), K. Baruch (NRGene) and S. Taudien (Universitätsmedizin Göttingen) for providing technical, laboratory, greenhouse or bioinformatics services. A. Fiebig, D. Arend, J. Bauernfeind, T. Münch and H. Miehe (IPK) provided IT services. We thank A. Graner for helpful advice. Research for this project was supported by funding from: the Czech Science Foundation (grant no. 17-17564S) to H.S.; the Agriculture and Agri-Food Canada International Collaboration Agri-Innovation Program to A.L.; the Natural Resources Institute Finland Innofood Stategic Funds program to A.S.; the Biotechnology and Biological Sciences Research Council Designing Future Wheat program (grant no. BB/P016855/1) to A. Hall; the German Federal Ministry of Education and Research (BMBF) to K.F.X.M. and U.S. (project de.NBI no. FKZ 031A536) and E.B. (project RYE-SELECT no. FKZ 0315946A); the German Federal Ministry of Food and Agriculture (BMEL) (WHEATSEQ no. 2819103915) to K.F.X.M.; HYBRO Saatzucht GmbH & Co. KG to D.S.; the European Regional Development Fund's plants as a tool for sustainable global development project (grant no. CZ.02.1.01/0.0/0.0/16_019/0000827) to J.D.; the 2Blades Foundation to B.W.; the Julius Kühn-Institute to B.H. and F.O.; KWS SAAT SE & Co. KGaA to V.K.; the Deutsche Forschungsgemeinschaft (grant

no. HO 1779/30-1) to A. Houben; the Montana Wheat and Barley Committee to H.B.; the Noble Research Institute, LLC to X.-F. M.; the Australian Research Council (grant no. CE140100008) to I.S. and J.M.; Genome Canada and Genome Prairie (grant no. CTAG2) to C.J.P.; the National Research Council Canada's Wheat Flagship Program to D.K. and A.S.; the Province of Saskatchewan Agriculture Development Fund to D.B.F.; the Bundesamt für Landwirtschaft, Bern (grant no. PGREL NN-0036) to B.K.; the Polish National Science Centre (grant nos. DEC-2015/19/B/NZ9/00921, DEC-2014/14/E/NZ9/00285 and 2015/17/B/NZ9/01694) to M.R.-T., H.B.-B., S.S. and B.M.

Author contributions

N.S. conceived the study and coordinated the research together with H.B., H.B.-B., B.H., A. Houben, J.J., V.K., B.K., J.L., A.L., K.F.X.M., M.M., D.M., X.M., H.Ö., F.O., C.J.P., J.P., N.R., A.H.S., U.S., S.S., V.T., M.R.-T. and B.B.H.W. M.T.R.-W. and M.M. carried out data generation and analysis for genome assembly and data integration. A.B. (*Secale* diversity panel), V.K. ('Lo7'), B.B., D.B.F., B.H., Q.L., C.J.P. ('Norstar' and 'Puma') and H.B.-B., B.H., V.K., B.M., S.S. and M.R.-T. (*Secale* genome size estimation panel) contributed to provision, curation, cultivation and phenotyping of genetic resources. B.B., A. Himmelbach, D.K., S.P., C.J.P. and A.S. generated the sequencing data. J.Č., J.D. and J.V. carried out the genome size estimation and chromosome flow sorting. E.B., H.Š. and H.T. produced the Bionano optical map. M.B. and A. Houben carried out the FISH analysis. A. Hall, G.K., J.K., T.L., K.F.X.M., D.S. and M. Spannagl contributed to the gene annotation. H.G., K.F.X.M., M. Spannagl and T.W. contributed to the repetitive element annotation and analysis. H.B. and B.S. contributed to the miRNA annotation. M.T.R.-W., M.M., M. Schreiber, H.S. and U.S. carried out the diversity and evolutionary analysis. M.T.R.-W. and M.M. carried out the Hi-C-based SV detection. B.H., M.H., B.K., C.P., B.S., N.T., A.V.V., B.B.H.W. and T.W. carried out the resistance gene identification and analysis. B.H., J.M. and I.S. carried out the fertility restorer gene prediction and analysis. L.G., M.M., J.P., M.T.R.-W. and B.H. carried out the wheat–rye introgression analysis. B.B., A.F., C.J.P. and S.W. carried out the low-temperature tolerance analysis. The manuscript was written by M.T.R.-W., B.H. and N.S., with input from all authors. All authors have read and approved the manuscript.

Competing interests

V.K. is an employee of KWS SAAT SE & Co. KGaA. D.S. is an employee of HYBRO Saatzucht GmbH & Co. KG. All other authors declare no competing interests.

Additional information

Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s41588-021-00807-0.

Correspondence and requests for materials should be addressed to N.S.

Peer review information *Nature Genetics* thanks Peter Morrell and the other, anonymous, reviewer(s) for their contribution to the peer review of this work.

Reprints and permissions information is available at www.nature.com/reprints.

nature research

Corresponding author(s): Nils Stein

Last updated by author(s): Dec 4, 2020

Reporting Summary

Nature Research wishes to improve the reproducibility of the work that we publish. This form provides structure for consistency and transparency in reporting. For further information on Nature Research policies, see our Editorial Policies and the Editorial Policy Checklist.

Statistics

For all statistical analyses, confirm that the following items are present in the figure legend, table legend, main text, or Methods section.

n/a	Confirmed
☐	☒ The exact sample size (*n*) for each experimental group/condition, given as a discrete number and unit of measurement
☐	☒ A statement on whether measurements were taken from distinct samples or whether the same sample was measured repeatedly
☐	☒ The statistical test(s) used AND whether they are one- or two-sided *Only common tests should be described solely by name; describe more complex techniques in the Methods section.*
☐	☒ A description of all covariates tested
☐	☒ A description of any assumptions or corrections, such as tests of normality and adjustment for multiple comparisons
☐	☒ A full description of the statistical parameters including central tendency (e.g. means) or other basic estimates (e.g. regression coefficient) AND variation (e.g. standard deviation) or associated estimates of uncertainty (e.g. confidence intervals)
☐	☒ For null hypothesis testing, the test statistic (e.g. *F*, *t*, *r*) with confidence intervals, effect sizes, degrees of freedom and *P* value noted *Give P values as exact values whenever suitable.*
☒	☐ For Bayesian analysis, information on the choice of priors and Markov chain Monte Carlo settings
☐	☒ For hierarchical and complex designs, identification of the appropriate level for tests and full reporting of outcomes
☐	☒ Estimates of effect sizes (e.g. Cohen's *d*, Pearson's *r*), indicating how they were calculated

Our web collection on statistics for biologists contains articles on many of the points above.

Software and code

Policy information about availability of computer code

Data collection	The majority of software used in the study was unambiguously used in data 'analysis', but some software could be said to overlap with data 'collection' (e.g. adapter trimming with cutadapt, or demultiplexing with splitgbs). For simplicity, we have listed all software used in the "Data Analysis" field.
Data analysis	Published software used are listed below. Specific usage details are given in the methods. 10X Loupe browser 2.1.1 10X LongRanger v2.2.0 AHRD v1.6 Augustus v3.3.2 bbduk v37.28 bcftools v1.9 Bionano Solve v3.1 Blast2GO software v5.2 bwa v0.7 cn.mops 1.12.0 CNVnator 0.4 Cuffcompare v2.2.1 cutadapt v1.9.1 DeNovoMagic3.0 assembly pipeline (proprietary, NRGene Israel) DESeq2 v3.11 dotter v4.22 EIGENSOFT v6.0.1 EMBOSS package v6.6 (Incl. MUSCLE, WATER, and ClustalW, GetORF)

EvidenceModeller v1.1.1
GeMoMa v1.5.3
GenomeThreader v1.7.1
GMAP v2017-01-14
Hisat2 v2.1.0
Hisat2 v2.1.0
HMMER v3.2.1
hmmscan v3.2.1
htseq v11.1
minimap2 v2.1
mmseqs2 Release 8-fac81
ncbi-blast-2.3.0+/2.3.1+/2.8.1+/2.9.0+
NCSS 97
NLR-Annotator Pipeline, last pulled September 2018
PacBio SMRTlink v5.1.0
RepeatMasker v4.0
samtools v1.9
SnpEff v4
Stringtie v1.3.6
TandemRepeatsFinder v4.09
Transdecoder v3.0.0
TRITEX Pipeline version corresponding to commit ID 2898e74, with minor function modifications and nontrivial plotting tasks available in a sourceable, annotated R function collection available at https://github.com/mtrw/Sc_genome_assembly
vmatch dbcluster v2.3.0

R v3.4.2
R packages (dependencies not listed; all updated including dependencies to their latest available CRAN/Bioconductor versions on 1 February 2018):
parallel
mhsmm
ASMap
SNPRelate
plyr
dplyr
magrittr
data.table
ggplot2
colorspace
zoo
stringi
igraph
ape
kernlab
e1071
lmerTest

In-house scripts:

get_alleles_at_position.c (https://github.com/mtrw/tim_bioinfo_tools/blob/master/blast_get_alleles_at_position.c)
blast_to_snps.c (https://github.com/mtrw/tim_bioinfo_tools/blob/master/blast_to_snps.c)
splitgbs.c (github.com/umngao/splitgbs)
SUmirFind.pl, SUmirFold.pl, SUmirScreen_v2.py, and SUmirLocate_v2.py, all available at https://github.com/hikmetbudak/miRNA-annotation

For manuscripts utilizing custom algorithms or software that are central to the research but not yet described in published literature, software must be made available to editors and reviewers. We strongly encourage code deposition in a community repository (e.g. GitHub). See the Nature Research guidelines for submitting code & software for further information.

Data

Policy information about availability of data

All manuscripts must include a data availability statement. This statement should provide the following information, where applicable:

- Accession codes, unique identifiers, or web links for publicly available datasets
- A list of figures that have associated raw data
- A description of any restrictions on data availability

The 'Lo7' assembly and gene feature annotation data are available via e!DAL with DOIs 10.5447/ipk/2020/33 and 10.5447/ipk/2020/29. The visual suite of resources for assembly assessment are stored at 10.5447/ipk/2020/32. Raw sequence data generated in the course of the study are available at ENA with accession numbers PRJEB32636 (PE and MP data for assembly), PRJEB32574 and PRJEB34626 (Hi-C), PRJEB34439 (10X), PRJEB32587 (CSS), PRJEB35392 (GBS data), and PRJEB35461 (RNAseq and IsoSeq for annotation of 'Lo7'). Chromium 10X and RNAseq data for 'Puma' and 'Norstar' are available at PRJNA564622. The SNP matrix used for rye population genetic analyses is available via e!DAL with DOI 10.5447/ipk/2020/31. GBS and sequence data generated for the USDA and CIMMYT wheat diversity panels are available at ENA with accession numbers PRJNA566410, PRJNA566408, and PRJNA566409. Optical map data and alignments are available at via e!DAL with DOI 10.5447/ipk/2020/30. High-stringency transposable element annotations (used for evolutionary analyses) are given in Supplementary Table 10, while the larger, low-stringency annotations (used for assembly quality comparisons) are available via e!DAL with DOI 10.5447/ipk/2020/34.

Field-specific reporting

Please select the one below that is the best fit for your research. If you are not sure, read the appropriate sections before making your selection.

☒ Life sciences ☐ Behavioural & social sciences ☐ Ecological, evolutionary & environmental sciences

For a reference copy of the document with all sections, see nature.com/documents/nr-reporting-summary-flat.pdf

Life sciences study design

All studies must disclose on these points even when the disclosure is negative.

Sample size	Various, for the different studies described in the manuscript. The effects of introgression of yield analysis involves 19,702 observations of 2,164 genotypes over 78 sites and 26 years. The flow cytometry was performed on five instances of each sample, each measured three times on different days. Gene expression profiling sequencing and cold hardiness of Norstar wheat and Puma rye lines were each measured along 12 time stages, the RNA being sampled from two plants at each. LT50 was measured in five individuals in each of three experimental replicate groups for each of for each of five pre-selected test temperatures for each line at each time point. Extended details are given in the manuscript materials. An all studies reported in this paper, sample sizes were maximised within the constrains of practicability, e.g. availability of germplasms/ genotypes, cost of sequencing, availability of field and greenhouse space etc.
Data exclusions	No data were excluded from the analyses reported.
Replication	No whole-experiment replication (of the kind that is standard in classical clinical trial design) was conducted in the agricultural, cytogenetic, and in silico studies described in this paper.
Randomization	Of the studies reported, randomisation is relevant only to the study on cold acclimation in rye and wheat plants. Norstar and Puma plants sampled for the expression and cold acclimation studies were grown in a randomised complete block design, with the three replicates separated in time and space between blocks.
Blinding	No blinding (of the kind that is standard in classical clinical trial design) was conducted in the agricultural, cytogenetic, and in silico studies described in this paper.

Reporting for specific materials, systems and methods

We require information from authors about some types of materials, experimental systems and methods used in many studies. Here, indicate whether each material, system or method listed is relevant to your study. If you are not sure if a list item applies to your research, read the appropriate section before selecting a response.

Materials & experimental systems

n/a	Involved in the study
☒	☐ Antibodies
☒	☐ Eukaryotic cell lines
☒	☐ Palaeontology and archaeology
☒	☐ Animals and other organisms
☒	☐ Human research participants
☒	☐ Clinical data
☒	☐ Dual use research of concern

Methods

n/a	Involved in the study
☒	☐ ChIP-seq
☐	☒ Flow cytometry
☒	☐ MRI-based neuroimaging

Flow Cytometry

Plots

Confirm that:

☐ The axis labels state the marker and fluorochrome used (e.g. CD4-FITC).

☐ The axis scales are clearly visible. Include numbers along axes only for bottom left plot of group (a 'group' is an analysis of identical markers).

☐ All plots are contour plots with outliers or pseudocolor plots.

☐ A numerical value for number of cells or percentage (with statistics) is provided.

Methodology

Sample preparation	Grains from fifteen diverse rye accessions were provided by nine providers listed in the supplementary tables. Plants of pea served as an internal reference standard in flow cytometric estimation of nuclear DNA content in all accessions, except of the tetraploid accession ACE-1, for which S. cereale line 'Lo7' was used as a reference. Seeds of pea (Pisum sativum cv. Ctirad) were obtained from Semo (Smržice, Czech Republic) breeding station. Plants were raised in garden compost in pots and maintained in a greenhouse until they reached a height of 10–20 cm. Ten mg of fresh leaf tissue of each of the rye accessions and the reference standard were chopped together in a 1 mL volume of LB01 solution2 using a razor blade. The resulting homogenate was filtered through a 50 µm nylon mesh. The filtrate was made up to 50 µg/mL propidium iodide and 50 µg/mL RNase.
Instrument	CyFlow Space flow cytometer (Sysmex Partec GmbH, Görlitz, Germany) equipped with a 532 nm green laser.
Software	The NCSS 97 statistical software package (Statistical Solutions Ltd.)
Cell population abundance	N/A. The gain of the instrument was adjusted so that the peak representing G1 nuclei of the genome size standard was positioned approximately on channel 100 on a histogram of relative fluorescence intensity when using a 512-channel scale.
Gating strategy	N/A, since the aim of the experiment was not to separate cell types. No plots or figures are included in association with the genome size estimation by flow cytometry. Data are available in the supplementary tables S1.

☐ Tick this box to confirm that a figure exemplifying the gating strategy is provided in the Supplementary Information.